# Rethinking Zero-shot Neural Machine Translation: From a Perspective of Latent Variables

**Weizhi Wang**[1]*, **Zhirui Zhang**[2]†, **Yichao Du**[3], **Boxing Chen**[2], **Jun Xie**[2], **and Weihua Luo**[2]
[1]Rutgers University, New Brunswick, USA
[2]Machine Intelligence Technology Lab, Alibaba DAMO Academy
[3]University of Science and Technology of China, China
[1]weizhi.wang@rutgers.edu [2]zrustc11@gmail.com
[2]{boxing.cbx, qingjing.xj, weihua.luowh}@alibaba-inc.com
[3]duyichao@mail.ustc.edu.cn

## Abstract

Zero-shot translation, directly translating between language pairs unseen in training, is a promising capability of multilingual neural machine translation (NMT). However, it usually suffers from capturing spurious correlations between the output language and language invariant semantics due to the maximum likelihood training objective, leading to poor transfer performance on zero-shot translation. In this paper, we introduce a denoising autoencoder objective based on pivot language into traditional training objective to improve the translation accuracy on zero-shot directions. The theoretical analysis from the perspective of latent variables shows that our approach actually implicitly maximizes the probability distributions for zero-shot directions. On two benchmark machine translation datasets, we demonstrate that the proposed method is able to effectively eliminate the spurious correlations and significantly outperforms state-of-the-art methods with a remarkable performance. Our code is available at https://github.com/Victorwz/zs-nmt-dae.

## 1 Introduction

Multilingual neural machine translation (NMT) system concatenates multiple language pairs into one single neural-based model, enabling translation on multiple language directions (Firat et al., 2016; Ha et al., 2016; Johnson et al., 2017; Kudugunta et al., 2019; Arivazhagan et al., 2019b; Zhang et al., 2020). Besides, the multilingual NMT system can achieve translation on unseen language pairs in training, and we refer to this setting as zero-shot NMT. This finding is promising that zero-shot translation halves the decoding time of pivot-based method and avoids the problem of error propagation. Meanwhile, zero-shot NMT casts off the requirement of parallel data for a potentially quadratic number of language pairs, which is sometimes impractical especially between low-resource languages. Despite the potential benefits, achieving high-quality zero-shot translation is a very challenging task. Standard multilingual NMT systems are sensitive to hyper-parameter settings and tend to generate poor outputs.

| Model | BLEU on DE⇒FR |
|---|---|
| DE⇒EN+EN⇒FR | 6.0 |
| PIV-(DE⇒EN+EN⇒FR) | 31.7 |

Table 1: BLEU scores [%] of training multilingual NMT with these two translation directions and its pivoting variant on Europarl Dataset.

One line of research believes that the success of zero-shot translation depends on the ability of the model to learn language invariant features, or an interlingua, for cross-lingual transfer (Arivazhagan et al., 2019a; Ji et al., 2020; Liu et al., 2021). Arivazhagan et al. (2019a) design auxiliary losses on the NMT encoder that impose representational invariance across languages. Ji et al. (2020) build up a universal encoder for different languages via bridge language model pre-training, while Liu et al. (2021) disentangle positional information in multilingual NMT to obtain language-agnostic representations. Besides, Gu et al. (2019) point out that the conventional multilingual NMT model heavily captures spurious correlations between the output language and language invariant semantics due to the maximum likelihood training objective, making it hard to generate a reasonable translation in an unseen language. Then they investigate the effectiveness of decoder pre-training and back-translation on this problem.

In this paper, we focus on English-centric multilingual NMT and propose to incorporate a simple denoising autoencoder objective based on English language into the traditional training objective of

*Contribution during internship at Alibaba.
†Corresponding author.



*Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4321–4327
November 7–11, 2021. 

multilingual NMT to achieve better performance on zero-shot directions. This approach is motivated by an observation that: as shown in Table 1, if we only optimize two translation directions DE⇒EN and EN⇒FR in a single model, it hardly achieves successful zero-shot translation on DE⇒FR. It is because that the model easily learns high mutual information between language semantics of German and output language, ignoring the functionality of language IDs. Actually, this mutual information can be significantly alleviated by directly replacing the original German sentence with a noisy target English sentence in training data, thereby guiding the model to learn the correct mapping between language IDs and output language. Besides, we analyze our proposed method by treating pivot language as latent variables and find that our approach actually implicitly maximizes the probability distributions for zero-shot translation directions.

We evaluate the proposed method on two public multilingual datasets with several English-centric language-pairs, Europarl (Koehn, 2005) and MultiUN (Ziemski et al., 2016). Experimental results demonstrate that our proposed method not only achieves significant improvement over vanilla multilingual NMT on zero-shot directions, but also outperforms previous state-of-the-art methods.

## 2 Multilingual NMT

The multilingual NMT system (Johnson et al., 2017) combines different language directions into one single translation model. Due to data limitations of non-English languages, multilingual NMT systems are mostly trained on large-scale English-centric corpus via maximizing the likelihood over all available language pairs $\mathcal{S}$:

$$\mathcal{L}_m(\theta) = \sum_{(i,j)\in\mathcal{S},(x,y)\in D^{i,j}} \log P(y|x,\mathbf{j};\theta), \quad (1)$$

where $(i,j) \in \mathcal{S}$ are the sampled source language ID and target language ID in all available language pairs, $D^{i,j}$ represents for the corresponding parallel data, and $\theta$ is the model parameter. The target language ID is appended as the initial token of source sentences, to let the model know which language it should translate to. In addition, the multilingual NMT system has proven the capability of translating on unseen pairs in training (Firat et al., 2016; Johnson et al., 2017), which is a property of **zero-shot translation**. However, the zero-shot translation quality significantly falls behind that of pivoting methods. The main issue leading to the unsatisfactory performance is that the multilingual NMT model captures spurious correlations between the output language and language invariant semantics due to the maximum likelihood training objective (Gu et al., 2019).

## 3 Method

In this section, we first introduce the denoising autoencoder task and then analyze the effectiveness of our proposed method from the perspective of latent variables.

**Denoising Autoencoder Task.** Given English-centric parallel data (X/Y/...⇔EN), we usually optimize the maximum likelihood training objective to build the multilingual NMT model. Since the target language ID is inserted at the beginning of the source sentence and only treated as a single token, the maximum likelihood training objective easily ignores the functionality of target language ID, leading to unreasonable mutual information between language semantic of "X/Y/..." and output language of English. To address this problem, we introduce a denoising sequence-to-sequence task, in which we directly replace the original input sentence with a noisy target English sentence in training data. In this way, previous mutual information can be significantly reduced, while enhancing the relationship between language IDs and output language. Specifically, we simply use all English sentences in parallel data to construct the denoising English corpus $D_{\text{EN}}$ via text infilling operation (Lewis et al., 2020). Then we optimize the multilingual NMT model via maximizing the original translation objective $\mathcal{L}_m(\theta)$ and denoising autoencoder objective $\mathcal{L}_d(\theta)$:

$$\mathcal{L}_d(\theta) = \sum_{j=\text{<2en>},(\overline{y},y)\in D_{\text{EN}}} \log P(y|\overline{y},\mathbf{j};\theta), \quad (2)$$

$$\mathcal{L}_a(\theta) = \mathcal{L}_m(\theta) + \mathcal{L}_d(\theta). \quad (3)$$

**Latent Variable Perspective.** As for zero-shot translation, we actually aim at directly fitting the probability distribution between non-English languages "X/Y/..." in the unified multilingual NMT system. For convenience, we consider the probability distribution $P(Y|X;D^*)$ between two non-English languages over the ideal parallel training data $D^*$. In practice, it is difficult to obtain such training data $D^*$ for the model training. To handle this issue, we convert the task of maximizing

$P(Y|X;D^*)$ into optimizing three existing sub-tasks, by treating the English language as a latent variable $h$ and introducing the probability distribution $P(h|\overline{h})$ of denoising autoencoder task:

$$\begin{aligned}
P(Y|X;D^*) &= \sum_{(x,y)\in D^*} \log P(y|x) \\
&= \sum_{(x,y)\in D^*} \log \sum_h P(y|h,x)P(h|x) \\
&\approx \sum_{(x,y)\in D^*} \log \sum_h P(h|\overline{h}) \frac{P(y|h)P(h|x)}{P(h|\overline{h})} \\
&\geq \sum_{(x,y)\in D^*} \sum_h P(h|\overline{h}) \log \frac{P(y|h)P(h|x)}{P(h|\overline{h})} \\
&= \sum_{(x,y)\in D^*} \mathbb{E}_{h\sim P(h|\overline{h})} \log P(y|h) \\
&\qquad - \mathbb{KL}(P(h|\overline{h})||P(h|x)) \\
&= P^*(Y|X;D^*,P(h|\overline{h})),
\end{aligned} \tag{4}$$

where we assume that $P(y|h,x) \approx P(y|h)$ due to the semantic equivalence of languages $h$ and $x$. With above equation, the original objective is transformed into optimizing three sub-tasks $P(h|x)$, $P(y|h)$ and $P(h|\overline{h})$. Incorporating the denoising autoencoder objective into the translation objective of multilingual NMT model helps minimize the KL-divergence terms, thus implicitly maximizing the lower bound of probability distributions of zero-shot directions. Following Ren et al. (2018), the gap between $P^*(Y|X;D^*,P(h|\overline{h}))$ and $P(Y|X;D^*)$ can be calculated as follow:

$$\begin{aligned}
\Delta &:= P(Y|X;D^*) - P^*(Y|X;D^*,P(h|\overline{h})) \\
&= \sum_{(x,y)\in D^*} \sum_h P(h|\overline{h}) \log \frac{P(h|\overline{h})P(y|x)}{P(y|h)P(h|x)} \\
&\approx \sum_{(x,y)\in D^*} \mathbb{KL}(P(h|\overline{h})||P(h|y)),
\end{aligned} \tag{5}$$

where we leverage an additional approximation that $P(h|x,y) \approx P(h|y)$ due to the semantic equivalence. Refer to Appendix A.2 for detailed derivations. Once we complement $P(h|y)$ into three sub-tasks mentioned before, this gap could be further reduced, resulting in better performance on zero-shot translation directions.

## 4 Experiments

### 4.1 Experimental Settings

**Datasets.** We evaluate the proposed method on two benchmark machine translation datasets, Europarl and MultiUN. The data statistics of two selected datasets are summarized in Table 2. BLEU (Papineni et al., 2002) is used as the metric for evaluating translation quality. For Europarl dataset, we select three European languages, Germany (De), French (Fr) and English (En). We remove all parallel sentences between De and Fr to ensure the zero-shot setting. We use WMT *devtest2006* as validation set and *test2006* as test set. For MultiUN, four languages are selected, Arabic (Ar), Chinese (Zh), Russian (Ru), and English (En). The selected languages are distributed in various language families, making the zero-shot language transfer more difficult. We use MultiUN standard validation and test sets to report the zero-shot performance. To differentiate language pairs, we follow Johnson et al. (2017) to append the language tag "<2Y>" on the source side for translating $X \Rightarrow Y$.

| Dataset | Language Pairs | Train | Dev & Test |
|---|---|---|---|
| Europarl | De-En, Fr-En | 1.8M | 2000 |
| MultiUN | Ar-En, Zh-En, Ru-En | 2M | 4000 |

Table 2: Data statistics of Europarl and MultiUN, in which we sub-sampled 2M samples for each language-pair in MultiUN.

**Baselines.** In our experiments, we compare the proposed method **MNMT+DN** with the following approaches: (*i*) **MNMT** (Johnson et al., 2017): training a multilingual NMT model on all directions with available parallel data; (*ii*) **LM+MNMT** (Gu et al., 2019): pre-training the decoder as a multilingual language model, then training the MNMT model initialized with the pre-trained decoder; (*iii*) **MNMT-RC** (Liu et al., 2021): removing residual connections in an encoder layer to disentangle positional information. We re-implement all baseline methods, following the same experimental settings to make fair comparison with our method.

**Experimental Details.** We choose standard Transformer-base (Vaswani et al., 2017) architecture to conduct experiments on all baseline and proposed methods, with $n_{\text{layer}} = 6, n_{\text{head}} = 8, d_{\text{embed}} = 512$. We use `faiseq` toolkit[1] (Ott et al., 2019) for fast implementations and experiments. We deploy Adam (Kingma and Ba, 2015) ($\beta_1 = 0.9, \beta_2 = 0.98$) optimizer and train all models with $lr = 0.0005, t_{\text{warmup}} = 4000$, dropout $=$

[1] https://github.com/pytorch/fairseq

| MultiUN | Ar, Zh, Ru ↔ En | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Model | Ar-Ru | | Ar-Zh | | Ru-Zh | | Zero Avg. | Parallel Avg. |
| | ← | → | ← | → | ← | → | | |
| MNMT | 17.9 | 13.4 | 16.1 | 29.5 | 12.1 | 30.3 | 19.9 | 49.2 |
| LM+MNMT | 22.0 | 29.3 | 20.3 | 42.7 | 24.3 | 42.1 | 30.1 | 48.9 |
| MNMT-RS | 20.8 | 26.1 | 20.3 | 37.9 | 24.2 | 37.4 | 27.8 | 49.9 |
| MNMT+DN (Ours) | **24.6** | **33.0** | **24.6** | **47.2** | **30.0** | **46.1** | **34.3** | **50.1** |

Table 3: Overall BLEU scores [%] on six zero-shot directions of MultiUN dataset. "Zero Avg." and "Parallel Avg." refer to average BLEU score of six zero-shot directions and six supervised directions, respectively.

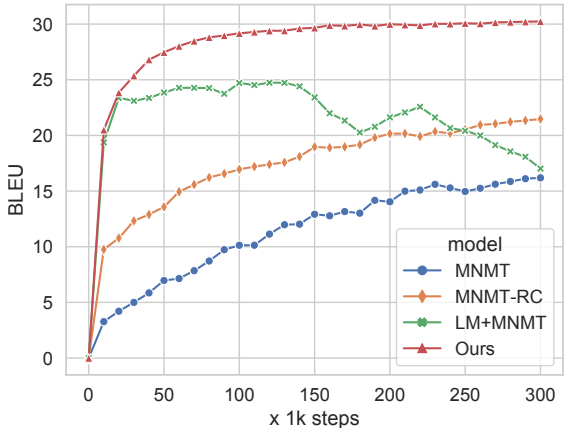


Figure 1: Learning curve of different methods on MultiUN dataset. We sub-sample 1K sentences from every zero-shot translation direction and report BLEU score on the combined 6K-size validation set.

| Europarl | De, Fr ↔ En | | | |
|---|---|---|---|---|
| Model | De-Fr | | Zero Avg. | Parallel Avg. |
| | ← | → | | |
| MNMT | 21.5 | 27.3 | 24.4 | 34.1 |
| LM+MNMT | 25.5 | 31.1 | 28.3 | 33.6 |
| MNMT-RC | 25.1 | 30.8 | 28.0 | 33.5 |
| MNMT+DN (Ours) | **27.1** | **31.8** | **29.5** | **33.7** |

Table 4: Overall BLEU scores [%] on two zero-shot directions of Europarl dataset.

$0.1, n_{\text{batch}} = 8000$ tokens. The `Moses` toolkit[2] (Koehn et al., 2007) is used to tokenize translation corpus. Exceptionally, we use Jieba[3] for Chinese tokenization. For each dataset, we lowercase all data and preprocess the corpus with 40K Byte-Pair-Encoding (BPE) (Sennrich et al., 2016) operations on all languages. For our proposed approach, we mask 30% of tokens in the whole training corpus, and deploy span masking (Joshi et al., 2020), in which a sequence text spans are sampled and masked, with the masked span lengths sampled from a Poisson distribution ($\lambda = 3$). 0-length spans correspond to the insertion of `[MASK]` token. Every model is trained for 300k updates on Europarl or 500K updates on MultiUN (additional 100k updates for pre-training), and the best model is selected based on BLEU score on validation set every 10k updates. For decoding, we adopt beam-search with beam size = 5 and calculate BLEU scores using `SacreBLEU`[4].

[2] https://github.com/moses-smt/mosesdecoder

[3] https://github.com/fxsjy/jieba

[4] https://github.com/mjpost/sacrebleu

### 4.2 Results on MultiUN Dataset

Table 3 reports the main results on the MultiUN dataset. We can find that our proposed method achieves state-of-the-art performance on all six zero-shot translation directions among all multilingual NMT systems. In addition, our method significantly improves the zero-shot performance of vanilla MNMT model by an average 14.4 BLEU score without performance degradation on supervised directions. These results demonstrate the effectiveness of incorporating denoising autoencoder objective in the training of multilingual NMT. We further investigate the learning curve of different methods on the validation set. As shown in Figure 1, our proposed method reaches faster convergence than MNMT and MNMT-RC, while LM+MNMT easily leads to over-fitting.

### 4.3 Results on Europarl Dataset

The main results on the Europarl dataset are presented in Table 4. We can observe that our proposed method still significantly improves the zero-shot translation performance of multilingual NMT systems with an average of 5.1 BLEU score improvements. Different from the MultiUN dataset with four languages distributed in different language families, the selected languages (De, Fr, En) of Europarl are all European languages, making the gap between various baselines and our method smaller than that of MultiUN.

| Model | Dataset | |
|---|---|---|
| | MultiUN | Europarl |
| MNMT | 57.49% | 98.19% |
| LM+MNMT | 91.87% | 99.13% |
| MNMT-RC | 83.37% | 99.00% |
| MNMT+DN (Ours) | **95.76%** | **99.13%** |

Table 5: Average language accuracy on all zero-shot directions of two selected datasets.

### 4.4 Evaluation of Off-Target Translations

We further summarize the percentage of off-target translations on zero-shot directions to verify the effectiveness of the proposed method. Generating off-target translations means that the multilingual NMT system fails in achieving zero-shot translation and generates translation in wrong output language. We use `langdetect`[5] toolkit to capture the off-target translations and calculate the language accuracy as $(1 - n_{\text{off-target}}/n_{\text{sentences}})$. The results of language accuracy on two selected corpora are presented in Table 5. The proposed method achieves the language accuracy of 99.13% on Europarl and 95.76% on MultiUN, which surpass baseline methods with a significant improvement. The results demonstrate that our method effectively alleviates the issue of off-target translation in zero-shot directions.

### 4.5 Ablation Study

As illustrated in Equation 4, the training objective of zero-shot directions can be converted into optimizing three sub-tasks jointly. To verify this analysis, we conduct an ablation study on the Europarl dataset. We consider a single model with two translation directions DE⇒EN+EN⇒FR. As shown in Table 6, when incorporating denoising autoencoder task, DE⇒EN+EN⇒FR+DN achieves a remarkable zero-shot performance on DE⇒FR of 31.1 BLEU score. This result demonstrates that the introduction of denoising autoencoder task can effectively break the spurious correlations between output language and semantics, enabling the failed model to perform zero-shot translation. Complementing with more translation tasks, such as FR⇒EN and EN⇒DE, MNMT+DN further improves translation accuracy on DE⇒FR, which proves the analysis of Equation 5. In addition, an alternative to our proposed method is BART pre-training (BART-PT), which first learns the denoising autoencoder objective and fine-tunes on the multilingual corpus. We can observe that BART-PT gains a similar performance to LM+MNMT, but worse than MNMT+DN due to the catastrophic forgetting problem (McCloskey and Cohen, 1989). The full results of BART-PT on MultiUN and Europarl datasets are illustrated in Appendix A.1.

[5] https://github.com/Mimino666/langdetect

| Europarl | De, Fr ↔ En | | | |
|---|---|---|---|---|
| Setting | De-Fr | | Zero | Parallel |
| | ← | → | Avg. | Avg. |
| DE⇒EN+EN⇒FR | - | 6.0 | - | - |
| DE⇒EN+EN⇒FR+DN | - | 31.1 | - | - |
| MNMT | 21.5 | 27.3 | 24.4 | 34.1 |
| BART-PT | 25.7 | 31.2 | 28.5 | 33.6 |
| MNMT+DN (Ours) | **27.1** | **31.8** | **29.5** | **33.7** |

Table 6: BLEU scores [%] of the ablation study on Europarl dataset. "+DN" means that the experiment setting includes denoising autoencoder task.

## 5 Conclusion

In this paper, we proposed to introduce denoising autoencoder objective into conventional translation objective to improve the zero-shot performance of multilingual NMT system. We analyze the motivation and effectiveness of proposed method from the perspective of latent variables. The experimental results demonstrate that our proposed method can significantly resolve spurious correlation issue in multilingual NMT and achieves state-of-the-art performance on zero-shot translation. In the future, it is interesting to explore the combination of our method and other language model pre-training methods (Song et al., 2019; Liu et al., 2020).

## Acknowledgment

We would like to thank the anonymous reviewers for the helpful comments. This work is supported by Alibaba Innovative Research Program. We appreciate Junliang Guo and Xin Zheng for the fruitful discussions. This work is done during the first author's internship at Alibaba DAMO Academy.

## References

N. Arivazhagan, Ankur Bapna, Orhan Firat, Roee Aharoni, M. Johnson, and Wolfgang Macherey. 2019a. The missing ingredient in zero-shot neural machine translation. *ArXiv*, abs/1903.07091.

N. Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, M. Johnson, M. Krikun, M. Chen, Yuan Cao, George F. Foster, Colin Cherry, Wolfgang Macherey, Z. Chen, and Y. Wu. 2019b. Massively

multilingual neural machine translation in the wild: Findings and challenges. *ArXiv*, abs/1907.05019.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *NAACL*.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and V. Li. 2019. Improved zero-shot neural machine translation via ignoring spurious correlations. In *ACL*.

Thanh-Le Ha, J. Niehues, and A. Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. *ArXiv*, abs/1611.04798.

Baijun Ji, Zhirui Zhang, Xiangyu Duan, Min Zhang, Boxing Chen, and Weihua Luo. 2020. Cross-lingual pre-training based transfer for zero-shot neural machine translation. In *AAAI*.

M. Johnson, M. Schuster, Quoc V. Le, M. Krikun, Yonghui Wu, Z. Chen, Nikhil Thorat, F. Viégas, M. Wattenberg, G. Corrado, Macduff Hughes, and J. Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL*.

Sneha Kudugunta, Ankur Bapna, Isaac Caswell, N. Arivazhagan, and Orhan Firat. 2019. Investigating multilingual nmt representations at scale. *ArXiv*, abs/1909.02197.

M. Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*.

Danni Liu, J. Niehues, James Cross, Francisco Guzmán, and Xian Li. 2021. Improving zero-shot translation by disentangling positional information. In *ACL/IJCNLP*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, M. Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

M. McCloskey and N. Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation*, 24:109–165.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, S. Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *NAACL*.

Kishore Papineni, S. Roukos, T. Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Shuo Ren, Wenhu Chen, Shujie Liu, Mu Li, M. Zhou, and S. Ma. 2018. Triangular architecture for rare language translation. *ArXiv*, abs/1805.04813.

Rico Sennrich, B. Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. *ArXiv*, abs/1508.07909.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. Mass: Masked sequence to sequence pre-training for language generation. In *ICML*.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *ArXiv*, abs/1706.03762.

Biao Zhang, P. Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *ACL*.

Michal Ziemski, Marcin Junczys-Dowmunt, and B. Pouliquen. 2016. The united nations parallel corpus v1.0. In *LREC*.

# A Appendix

## A.1 Full Results on MultiUN and Europarl

We report the performance of zero-shot and supervised translations on MultiUN and Europarl in Table 7 and 8. We also include the pivoting version of MNMT: PIV-M. Our proposed method still lags behind the pivoting method by an average BLEU score of 4.3 on MultiUN dataset, while achieving slightly better performance on Europarl dataset. Besides, our method outperforms BART pre-training by an average BLEU score of 2.6/1.0 on MultiUN and Europarl, respectively.

| MultiUN | Ar, Zh, Ru ↔ En | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Ar-Ru ← | Ar-Ru → | Ar-Zh ← | Ar-Zh → | Ru-Zh ← | Ru-Zh → | Zero Avg. | En-Ar ← | En-Ar → | En-Zh ← | En-Zh → | En-Ru ← | En-Ru → | Parallel Avg. |
| PIV-M | **29.9** | **36.8** | **29.2** | **51.5** | **34.3** | **50.1** | **38.6** | 54.7 | 37.8 | 50.7 | 58.3 | 50.7 | 42.8 | 49.2 |
| MNMT | 17.9 | 13.4 | 16.1 | 29.5 | 12.1 | 30.3 | 19.9 | 54.7 | 37.8 | 50.7 | 58.3 | 50.7 | 42.8 | 49.2 |
| LM+MNMT | 22.0 | 29.3 | 20.3 | 42.7 | 24.3 | 42.1 | 30.1 | 54.4 | 37.3 | 50.7 | 57.7 | 50.7 | 42.8 | 48.9 |
| MNMT-RS | 20.8 | 26.1 | 20.3 | 37.9 | 24.2 | 37.4 | 27.8 | 55.6 | 38.3 | 51.6 | 58.7 | 51.6 | 43.4 | 49.9 |
| BART-PT | 22.9 | 30.2 | 22.3 | 44.1 | 27.8 | 43.1 | 31.7 | 53.8 | 37.3 | 49.8 | 57.3 | 50.0 | 42.0 | 48.4 |
| MNMT+DN (Ours) | 24.6 | 33.0 | 24.6 | 47.2 | 30.0 | 46.1 | 34.3 | **56.1** | **38.0** | **52.1** | **58.6** | **52.0** | **43.9** | **50.1** |

Table 7: Overall BLEU scores [%] on six zero-shot directions and six supervised directions of MultiUN dataset. "Zero Avg." and "Parallel Avg." refer to average BLEU score of six zero-shot directions and six supervised directions, respectively.

| Europarl | De, Fr ↔ En | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Model | De-Fr ← | De-Fr → | Zero Avg. | En-De ← | En-De → | En-Fr ← | En-Fr → | Parallel Avg. |
| PIV-M | 26.5 | 31.7 | 29.1 | **34.3** | **27.8** | **37.2** | **37.0** | **34.1** |
| MNMT | 21.5 | 27.3 | 24.4 | **34.3** | **27.8** | **37.2** | **37.0** | **34.1** |
| LM+MNMT | 25.5 | 31.1 | 28.3 | 33.8 | 27.2 | 36.8 | 36.4 | 33.6 |
| MNMT-RC | 25.1 | 30.8 | 28.0 | 33.5 | 27.5 | 36.6 | 36.5 | 33.5 |
| BART-PT | 25.7 | 31.2 | 28.5 | 33.6 | 27.4 | 36.8 | 36.6 | 33.6 |
| MNMT+DN (Ours) | **27.1** | **31.8** | **29.5** | 33.8 | 27.5 | 36.8 | 36.7 | 33.7 |

Table 8: Overall BLEU scores [%] on two zero-shot directions and four supervised directions of Europarl dataset. "Zero Avg." and "Parallel Avg." refer to average BLEU score of two zero-shot directions and four supervised directions, respectively.

### A.2 Derivations for Equations

The detailed derivations for latent distribution $P^*(Y|X; D^*, P(h|\overline{h}))$ are shown in Equation 4, while the derivations for the probability gap $\Delta$ in Equation 5 are as follows:

$$
\begin{aligned}
&\Delta := P(Y|X; D^*) - P^*(Y|X; D^*, P(h|\overline{h})) \\
&= \sum_{(x,y)\in D^*} \sum_{h} P(h|\overline{h}) \log \frac{P(h|\overline{h})P(y|x)}{P(y|h)P(h|x)} \\
&= \sum_{(x,y)\in D^*} \sum_{h} P(h|\overline{h}) \log \frac{P(h|\overline{h})P(y|x)P(h|y)}{P(y|h)P(h|x)P(h|y)} \\
&\approx \sum_{(x,y)\in D^*} \sum_{h} P(h|\overline{h}) \log \frac{P(h|\overline{h})P(y|x)P(h|y)}{P(y|h,x)P(h|x)P(h|y)} \\
&= \sum_{(x,y)\in D^*} \sum_{h} P(h|\overline{h}) \log \frac{P(h|\overline{h})P(y|x)P(h|y)}{P(y,h|x)P(h|y)} \\
&= \sum_{(x,y)\in D^*} \sum_{h} P(h|\overline{h}) \log \frac{P(h|\overline{h})P(y|x)P(h|y)}{P(h|x,y)P(y|x)P(h|y)} \\
&\approx \sum_{(x,y)\in D^*} \sum_{h} P(h|\overline{h}) \log \frac{P(h|\overline{h})P(y|x)P(h|y)}{P(h|y)P(y|x)P(h|y)} \\
&= \sum_{(x,y)\in D^*} \sum_{h} P(h|\overline{h}) \log \frac{P(h|\overline{h})}{P(h|y)} \\
&= \sum_{(x,y)\in D^*} \mathbb{KL}(P(h|\overline{h})||P(h|y)),
\end{aligned}
$$

where we use two approximations here that are $P(y|h,x) \approx P(y|h)$ and $P(h|x,y) \approx P(h|y)$.